Buffered Multistage Networks with Dynamic Routing

Jon Turner
Computer Science & Engineering
Washington University

www.arl.wustl.edu/~jst
Buffered Multistage Networks

- In unicast switch input adds output port number
- *Switch elements* use output port to route packets
- With *dynamic routing*, first stage distributes traffic
  » potential for out-of-order delivery requires resequencing
- Various buffering, inter-stage flow control options
  » affects delay performance and throughput
Shared Buffer Switch Element

- **Datapath** stores cells in on-chip SRAM
  - mux presents cells to SRAM sequentially
  - demux routes to individual outputs
- **Controller** maintains linked list queues
- **Extensions**
  - separate queues for different priorities
  - multicast cell replication
Parallel Planes for High Throughput

- SEs with many “skinny” links can be more efficient
  - fewer stages, fewer components for given throughput
  - “upstream” signal for flow control can become issue
- May divide cells into pieces or distribute different cells to different planes
Folded Network Topologies

- *Folded network topologies* use bidirectional data paths
  - data flows from left to right then back
  - shortcut paths may be used (or not)
  - $k$ level folded network corresponds to $2k-1$ stage unfolded network
- With narrow datapaths, folded networks can scale better
  - piggyback flow control information
  - example: if technology limits chips to 64 inputs, 64 outputs, largest 2 level folded net has 2048 ports, largest 3 stage unfolded net has 1024 ports
- Natural choice for datacenter networks
Benes Network - Structure & Routing

- Network expanded by adding stages on left, right
  » 2^{k-1} stage net with d port SEs supports up to d^k ports
- First k-1 stages (of 2k-1) do traffic distribution
  » guarantees internal link load ≤ external link load
- Route using base d digits of output address
Simulation Results

- Five stage Benes network with 8 port SEs
  » shared buffer SEs
  » grant flow control
  » uniform, random traffic
  » delays small so long as load below critical point
  » resequencer accounts for most delay at low loads

- Note: no sustained overloads at outputs
  » require higher level traffic regulation
Handling Unpredictable Traffic

- Sustained local overloads cause poor performance
  - overloaded outputs cause congestion
  - can affect outputs that are not overloaded
- To cope with short-term overloads
  - limit buffer sharing; finer-grained queueing & flow control
- To cope with long-term overloads
  - regulate system-wide traffic flows to prevent congestion
    - assign each input a share of each output's bandwidth
    - assign shares based on amount/importance of traffic
  - uses rate-controlled Virtual Output Queues (VOQ)
  - inputs/outputs periodically exchange information and adjust rates (alternatively, could use central controller)
Resequencing & Load Distribution

- Resequencing needed to restore cell order
  - low delay variance in systems with well-regulated traffic
    - small output-side resequencing buffers work well
  - two main approaches: timestamps or sequence numbers
    - timestamp approach simpler and applies more easily to multicast, but adds more delay

- Alternative is for SEs to guarantee forwarding cells in timestamp order
  - requires forwarding of “future timestamp lower-bounds”
  - requires more complex control, larger queues

- Delay variation sensitive to how is load distributed
  - want each output’s load evenly distributed over all paths
  - per-output random permutations work well
Other Issues

- Expanding network in running systems
  - want to minimize changes to inter-rack cables
  - in 3 stage nets with stages 1,3 in line card racks
    - add line card racks as needed
    - add switching racks for center stage less often
      - for typical designs, each expansion doubles capacity
    - if center stage built out in advance, need never move cables

- Coping with failures
  - in large systems, link, SE failures are routine occurrence
  - various strategies
    - detect and track failed links, SEs and route around them
    - provide extra switch planes and use only fault-free planes
      - in systems with 10 or more planes, a few extra suffice
Data Center Networking Issues

- Data center networks larger than WAN routers
  - 10K-100K servers with 1-10 Gb/s per server
    - implies aggregate capacities from 10 Tb/s to 1 Pb/s
    - incremental growth, reliability issues even more important
- Can be assembled using commodity switches or purpose-built systems
  - commodity case requires solutions for traffic regulation, load-balancing, resequencing
  - may implement in host-driver or programmable NIC
  - alternate approach distributes traffic by flows
    - no resequencing required
    - rely on TCP congestion control to limit load in network
    - can produce poor load balance